THE DECTALK SYSTEM FOR GERMAN: A STUDY OF THE MODIFICATION OF A TEXT-TO-SPEECH CONVERTER FOR A FOREIGN LANGUAGE

Gerhard Rigoll †

Fraunhofer-Institute (IAO)
Dept. for Advanced Information and Communication Technologies
Holzgartenstr. 17, 7000 Stuttgart 1, West Germany

ABSTRACT

This paper describes the development of the German version of the DECtalk system, which was originally designed for the American language by D.H. Klatt. The aim of this paper is not only to provide an overview on the problems and difficulties for German text-to-speech conversion, using the cascade/parallel formant synthesizer and on the use of new algorithms for parameter extraction, but also to provide a study of the modification procedure which is necessary to build a new language version for a text-to-speech system which was designed for a different language. These experiences are important for the future design of multilingual text-to-speech systems because the modification from one language to another language gives automatically the answer to many questions which are interesting for the design of a multilingual system or at least a system which can be easily modified for another language. The paper describes the most important steps that have to be performed during the modification procedure, i.e. text normalization and letter-to-sound rules, description of the used synthesizer, description of the tools used for speech analysis and comparison of synthetic and natural speech, tuning of the stationary parts of the phonemes, transitions, prosodics and generation of different voices.

INTRODUCTION

The DECtalk text-to-speech converter was originally developed by D.H. Klatt for the American language and became one of the most successful text-to-speech systems. A few years ago, it was decided to develop foreign versions of the DECtalk system to cover the main European languages. One of the most advanced foreign versions is the German version which was started in 1984 [1]. The work on the German DECtalk system is now almost completed and in a final review it can be said that the modification of a text-to-speech system for a new language is a very time consuming task which requires a high amount of language specific knowledge in phonetics and linguistics and of knowledge in signal processing and program development. It turned out that all important modules of the software for the text-to-speech system which are shown in Fig.1 have to be modified or even completely replaced. The most important modifications are described in the following sections.

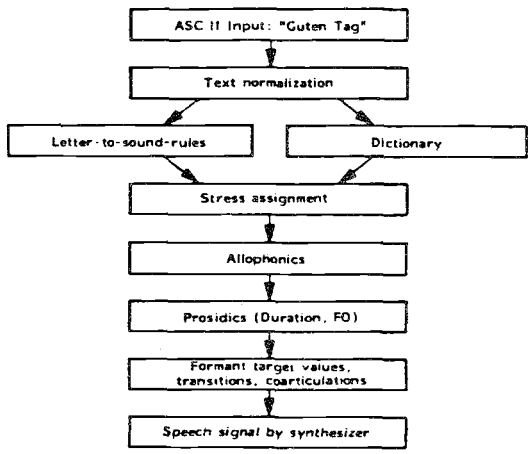

Fig.1: Basic modules of a text-to speech converter

TEXT NORMALIZATION AND LETTER-TO-SOUND RULES

The problems for text normalization and phonemization of orthographic text are generally not bigger in the German language than in the American language. While text normalization is relatively straightforward and can be done using mainly table lookup, the phonemization of German words is not quite uncomplicated because many long German words exist, consisting of several morphemes which are tied together with prefixes and suffixes. Therefore, the morphologic decomposition of a word is the most crucial step in the phonemization procedure. This can be done either by investigation of consonant clusters [4] which leads to a considerably high error rate or by the use of a morpheme dictionary which occupies a relatively large storage space. Once the morpheme boundaries are found, the phonemization of the particular morphemes is much easier than the phonemization of English morphemes. It is also obvious, that the correct stress assignment also depends on the correct detection of the morpheme boundaries. The fact that many foreign

† The author is a postdoctoral fellow in the Continuous Speech Recognition Group, Dept. of Computer Sciences, IBM Thomas J. Watson Research Center, Yorktown Heights, U.S.A., in 1987

1450 CH2396-0/87/0000-1450 $1.00

words exist in the German language makes the stress assignment even more complicated. Usually, the introduction of stress markers for primary and secondary stress is sufficient, but for very long words, a marker for tertiary stress may sometimes be appropriate.

THE CASCADE/PARALLEL FORMANT SYNTHESIZER

The cascade/parallel formant synthesizer used for the text-to-speech system is a slightly modified version of the synthesizer described by D.H. Klatt in [2].

One of the most remarkable changes is the new voicing source which is very flexible and is also important for the generation of different voices. The synthesizer has a large variety of control parameters, which gives the developer a high degree of freedom to synthesize all possible phonetic phenomena but this high degree of freedom leads sometimes also to the problem of finding the right parameter constellation to synthesize a desired utterance. The following is a short overview of the synthesizer parameters:

(1) parameters which are modified during the synthesis and which lead to different sounds:

- first 3 formants and bandwidths
- formant frequency of nasal anti-resonator
- gains of exitation sources
- gains of parallel resonators
- fundamental frequency

(2) parameters which are modified to obtain different speaker characteristics:

- formant and bandwidth of 4th and 5th resonator
- parameters which determine the shape of the voicing signal
- gains of the cascade resonators

(3) constant parameters:

- formant and bandwidth of 6th resonator
- coefficients of several low pass filters

TOOLS FOR SPEECH ANALYSIS AND SPECTRAL MATCHING OF NATURAL AND SYNTHETIC SPEECH

To obtain a high quality of the synthetic speech, it is necessary to study in detail the stationary and the transitional parts of all phonemes of the concerned language. For that purpose, software tools for formant and fundamental frequency analysis are required.

The spectrogram usually provides a more qualitative representation of the utterance which may be used for labeling the speech signal and for the investigation of longer utterances, if fine tuning was already performed. For the detailed quantitative spectral matching, the spectrum has to be used, which represents the speech signal only within a limited time window. Especially for the investigation of transitions it is sometimes useful to have a detailed information about the formant constellation over a longer time period. For this purpose some special formant tracking tools where developed which are based on a mathematical model of the cascade/parallel formant synthesizer. It was expected that an analysis on the base of this mathematical model would lead to values for the formants and bandwidths which lead to a good result, if these values are taken for a resynthesis of the analysed utterance.

In the mathematical model, the resulting synthesizer output is modeled directly as a function of the synthesizer parameters, e.g. the formants and the bandwidths. It turned out that the model for the cascade branch leads to the same formant tracking algorithm that was described by the author in [3], while a different model has to be used for the parallel branch. This model is based on the transfer function of the parallel branch:

$$G(z) = \frac{\sum_{i=1}^{m} \left[A_i \cdot \prod_{\substack{j=1 \\ j \neq i}}^{m} \left(1 + c_j \cdot z^{-1} + d_j z^{-2}\right) \right] + A_b \cdot \prod_{j=1}^{m} \left(1 + c_j z^{-1} + d_j z^{-2}\right)}{\prod_{i=1}^{m} \left(1 + c_i z^{-1} + d_i z^{-2}\right)}$$

where c_i and d_i are the coefficients of the digital resonators ([3]), A_i are the resonator gains (A_b bypass) and m is the number of resonators (i.e. m=6). A similar approach which is - as in [3] - based on Kalman-Filtering, can then be used to analyse obstruents which are supposed to be produced by the parallel branch. Fig.2 shows in the first part the spectrogram of the German diphtong /au/ which is difficult to analyse using a spectrogram because the first 2 formants are very close together. In the second part, the analysis with the Kalman-Filter is shown, using the model for the cascade branch.

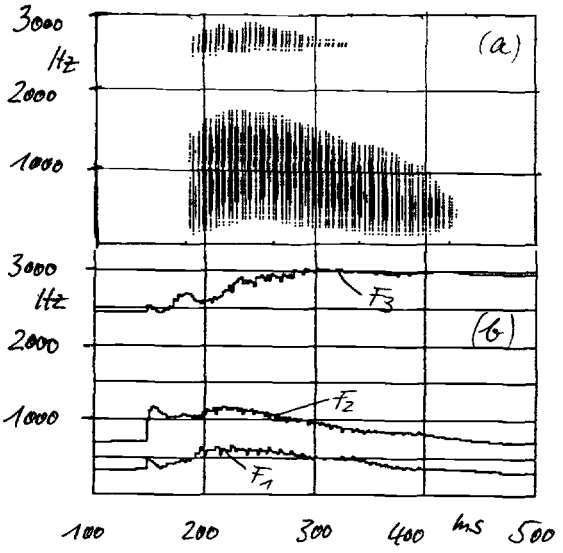

Fig.2: Formants of the German diphtong /au/ displayed with the use of a spectrogram (a) and with the use of a Kalman-Filter based formant tracker (b)

For obstruents, the filter tries to estimate the formants and the resonator gains from the given speech signal. The bandwidths

are not very important for the synthesis of these sounds and are not included in the state vector for the parallel branch model. This way of estimation sometimes still leads to problems for the analysis of obstruents but often yields good results. Fig.3 shows an example where the formant tracks for the generation of 2 voiceless fricatives were given to the parallel synthesizer branch and these tracks were recovered with a Kalman-Filter from the resulting synthetic speech signal with the use of the model for the parallel synthesizer branch.

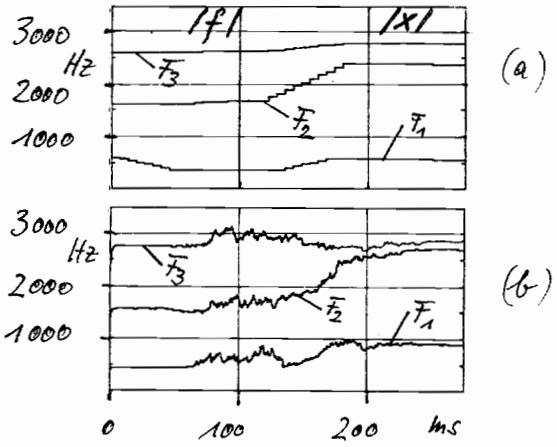

Fig.3: Formant tracks used for the synthesis of 2 German fricatives (a) and estimated formant tracks from the synthetic speech signal (b)

Since the synthesis with parameter values which were obtained from the formant analysis of an utterance, usually does not result in a speech signal, that has exactly the same spectral properties as the original speech signal, a time consuming tuning procedure is normally required to adapt the acoustic properties of the synthesized signal. The initial parameter values for this iteration can be obtained with the described tools, which can also be used for the spectral comparison of synthetic and natural utterance during the tuning process.

INVESTIGATION OF THE STATIONARY AND TRANSITIONAL PARTS OF THE PHONEMES

To obtain a high voice quality, it is necessary to perform spectral matching with all phonemes of the language and to investigate especially the consonant vowel transitions. The number of German phonemes is about the same as for the American language, including the same phoneme classes, like vowels, sonorants, stops, fricatives etc.. The consonant vowel transitions are calculated using consequently the locus theory, which is also the case for the American version. It turned out that the problems which occur with the use of that approach are also similar for both languages.

PROSODICS

For the calculation of the phoneme durations, the same basic model as for the American version [5], can be used for the German version. However the calculation of the context dependent lengthening or shortening rules is in some cases very similar for both languages (e.g. lengthening of stressed syllabic elements) and of course for phoneme specific rules completely different. This is also valid for the sentence intonation, where it turned out that the "hat pattern principle" used for the American intonation is completely inappropriate for the German synthesis. An intonation model based on the decomposition of a German sentence into syntactic sub-units was used [6]. This requires a rough syntactic analysis of the sentence, based on word classes. Fig.4 shows an example for the intonation of an American and a German sentence.

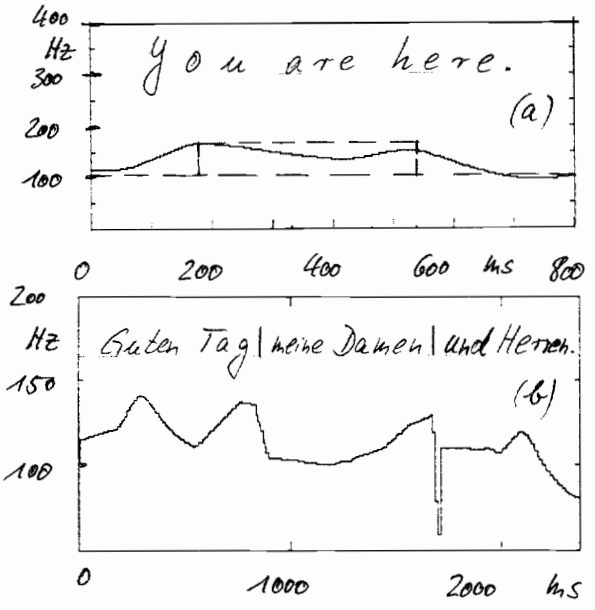

Fig.4: F0-contour for the intonation of a short American and a German sentence

GENERATION OF DIFFERENT VOICES

The capability of the American DECtalk system to generate different voices was also retained for the German version. Certain voice parameters are reserved for this purpose:

(1) excitation source parameters, which modify the shape of the voicing signal. Examples are the parameters for breathiness or laryngealization.

(2) vocal tract parameters as e.g. head size or the 4th formant modify the acoustic properties of the vocal tract transfer function.

(3) prosodic parameters as e.g. the average pitch or the pitch range are used to change the speaker sex or to modify the intonation of a speaker. Fig.5 shows the spectrum of the same vowel for a normal male voice and for a child voice.

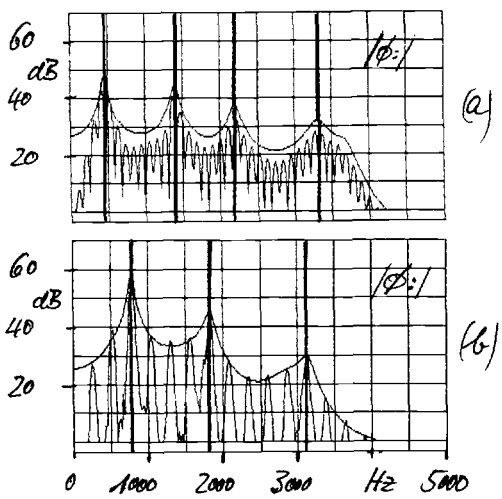

Fig.5: Spectrum of a vowel for a normal male voice (a) and a child voice (b)

SPECIAL PROBLEMS FOR THE GERMAN VERSION OF DECTALK

It is not possible to describe all problems that arose during the modification of the DECtalk system for the German language. The main points are very briefly listed below:

- letter-to-sound rules
- rules for stress assignment
- dictionary
- new duration rules
- syntactic analysis of the sentence
- new intonation rules
- new class of vowels (front-round)
- new fricatives (palatal and velar)
- sonorants /l/ and /ʔ/ without American accent
- synthesis of uvular /R/
- new German affricates
- nasalized vowels for German foreign words

CONCLUSION

During the development of the German version of the DECtalk text-to-speech system, many experiences were made, which main features a text-to-speech system should have in order to obtain high voice quality. These main features are summarized below:

- cascade/parallel formant synthesizer
- natural and very flexible voicing source
- complicated transition rules
- context depended duration rules; duration can be adjusted continuously
- complicated syntactic analysis of the sentence; consideration of many exceptions (e.g. stressed vowel at the end of the sentence etc.)
- complicated calculation of fundamental frequency; f0 can be adjusted continuously for each frame.
- large variability of all synthesizer parameters for the generation of different voices

REFERENCES

[1] Rigoll, G.: Development of the German Version of an American Text-to-Speech Converter. Proc. Speech Tech, New York 1985.

[2] Klatt, D.H.: Software for a cascade/parallel formant synthesizer. J.A.S.A., Vol. 67, No.3, 1980.

[3] Rigoll, G.: A New Algorithm for Estimation of Formant Trajectories Directly from the Speech Signal Based on an Extended Kalman Filter. Proc. IEEE-ICASSP, Tokyo 1986.

[4] Ruehl, H.-W.: SYNTEX - A Microprocessor based System for Automatic Conversion of German Text to Speech. Proc. IEEE-ICASSP, Paris 1982.

[5] Klatt, D.H.: Synthesis by Rule of Segmental Durations in English Sentences. Proc. 9th Int. Congress of Phonetic Sciences, Copenhagen 1979.

[6] Wolf, H.E.: Control of Prosodic Parameters for a Formant Synthesizer based on Diphone Concatenation. Proc. IEEE-ICASSP, Atlanta 1981.